The Data Acquisition, Accessibility, Annotation and e-Research Technologies (DART) Project: Supporting the complete e-Research Lifecycle

Andrew Treloar

{DART Technical Architect | ARCHER Director and Chief Architect | ARROW Technical Architect}, Monash University

Clayton Campus, Clayton, Victoria 3168, Australia

Abstract

The DART Project undertook a coordinated program of e-Research requirements analysis, software development, policy and guideline creation and prototyping to investigate how best to:

- collect, capture and retain large data sets and streams from a range of different sources;
- deal with the infrastructural issues of scale, sustainability and interoperability between repositories;
- support deposit into, access to, and annotation by a range of actors, to a set of digital libraries which include publications, datasets, simulations (data and model), software and dynamic knowledge representations;
- assist researchers in dealing with intellectual property issues during the research process;
- adopt next-generation methods for research publication, dissemination and access.

The DART project completed in June 2007. This paper presents an overview of the project, describes some of the issues encountered and lessons learned in solving some of these goals, and describes further work now underway.

1. Introduction

1.1 Background

In order to meet the emerging needs of e-Researchers, the Data Acquisition, Accessibility, Annotation in e-Research Technologies (DART) project undertook a comprehensive end-to-end approach to developing a new system for managing research activity and publication. The intention was that demonstrating the value of this way of managing the research process would have a far-reaching impact on researcher practice.

The DART project sought to respond to rapid and ongoing changes in the way research is carried out, and the way results are communicated. In formulating the project, its creators identified a number of key trends that are changing the way research is conducted and its outputs consumed. These include:

- new technologies, such as extensive computer simulations, availability of large data sets from instruments such as synchrotrons and sensor networks;
- the rapidly expanding size of data sets on which research is based;
- increasing volumes of information generated through research;
- greater complexity in the conduct of research, in that researchers wish to collaborate with other researchers, who may be located elsewhere;
- recognition of the need to work across traditional disciplinary, institutional and national borders; and,
- a growth in research practices that are producing a paradigm shift in the types of research that this new large-scale computing/data management environment can support.

These emerging research practices:

- are intensely collaborative (often involving trans-national teams);
- require high-quality network access; and,
- are data and simulation-intensive.

These changes first became evident in high-energy physics, science and engineering [1] but are now also becoming apparent in the social sciences and humanities [2].

In 2005, a joint CNI-JISC-SURF invitational conference was held in Amsterdam with the title

"Making the strategic case for institutional repositories". This conference emphasized the potential for repositories to move beyond the kinds of traditional publications that have been the concern of the open access movement tosupport innovative new forms of research and research output exposure. Some of the possibilities discussed were:

- life cycle management of research (from laboratory log book to formal outputs to teaching);
- smart publications that link experiments, results, and a range of supporting documents;
- the ability to validate not only research conclusions, but also research results, by replication and comparison;
- the ability to allow other researchers access to original raw data which might result in quite different discoveries and possibly more important discoveries by someone other than the generator of the original research (a form of post processing and knowledge mining);
- the potential to shorten the "publication cycle" (time to release information about new research);
- environments that provide stronger support for authenticity, authority, and integrity of research.

All of these new possibilities, of course, also present new challenges in life-cycle management, attribution and provenance of the full set of research outputs, not just the conventional formal publication [14].

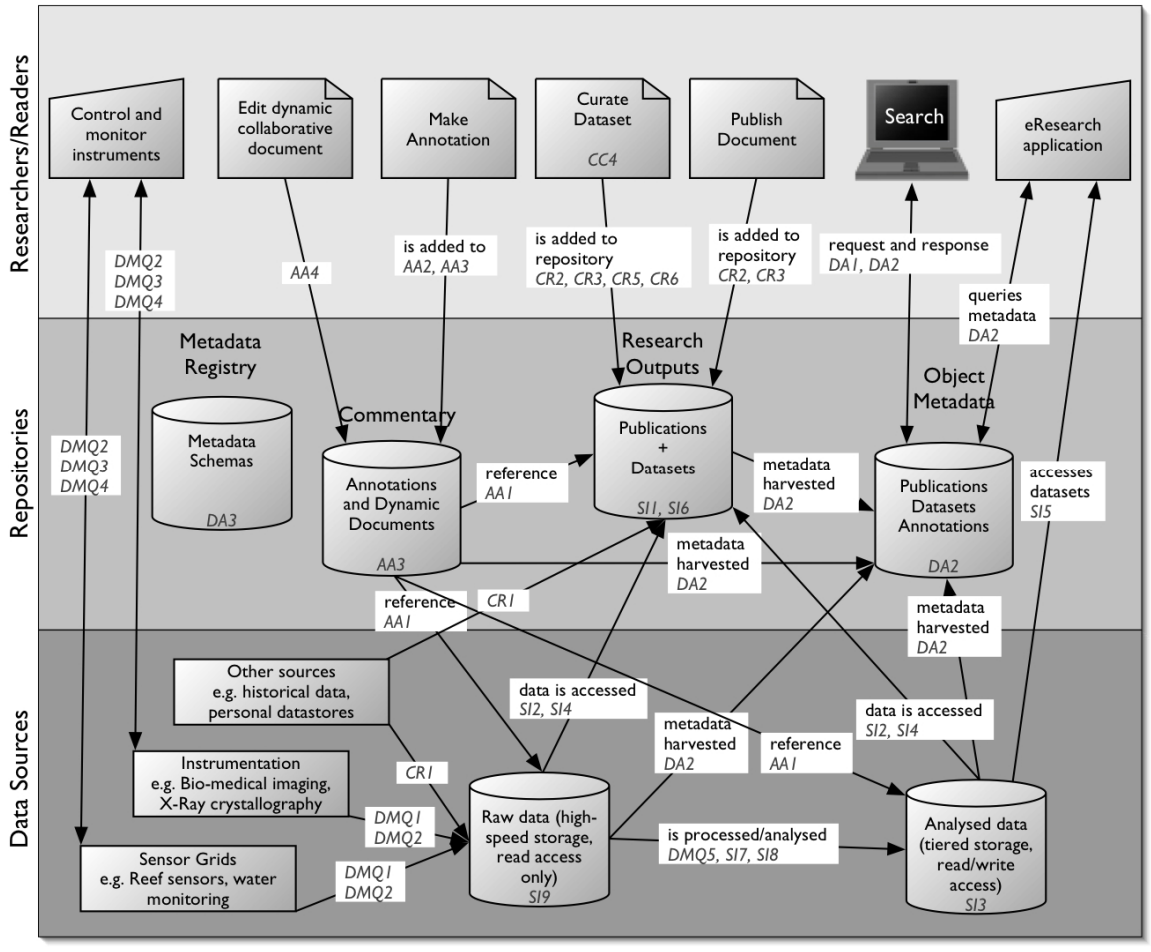

Figure 1: DART high-level architecture

1.2 Inception

The DART project came into being around the middle of 2005 in response to a call for projects from the Australian Commonwealth Department of Education, Science and Training. The project was a collaboration between Monash University (lead institution), James Cook University and the University of Queensland. The project was successful in receiving the requested funding of A$3.23M in August 2005 and the project formally commenced in December 2005 with a notional sunset of December 2006 (later extended to June 2007).

1.4 Objectives

The specific objectives of the DART project were to demonstrate the value of a system that could:

support and enable researchers, end-users, and appropriate computer systems to manage the creation and collection of data and to gain greater access to data and documents by *gathering, managing and archiving data and documents* and managing their access so that *researchers were more easily able to perform their work* and do so at a much higher level of insight and productivity than was previously possible, and so that the *Australian public had greater visibility of, and access to, publicly funded research.*

It is important to note that DART was always intended as proof-of-concept project. DART was never intended to provide production services. This is being taken up in the successor ARCHER project discussed later.

2. Project Architecture

The DART high-level architecture is shown in Figure 1. In the uppermost layer are researchers, readers and computer programs. The middle layer shows the repositories (including traditional publications as research outputs, and raw data) and the data flows between them and the datasets in the lowest layer. The lowest layer shows the data sources and their associated storage.

The figure has been annotated to indicate which work packages are involved for each component (more details about the DART workpackages are available on the DART website at http://dart.edu.au/. and an overview is provided in the next section). The project thus substantially extended and enhanced the focus on research outputs to include the needs of dataset creation, acquisition, management, and curation, as well as providing support for collaborative research practices.

3. Project Overview

The DART project was structured as a number of inter-related thematically-grouped sets of work packages. For each set of work packages, an overview of the functionality developed, and the rationale for its inclusion are described below.

3.1 Data Collection, Monitoring and Quality Assurance

In this group of work packages, DART tackled the issues surrounding high-rate and large-volume data streams, particularly those generated by instruments and sensors. There are a number of requirements that are unique to the challenges inherent in dealing with digital objects generated by and derived from instruments and sensors:

- Two way communication with the instruments and sensors so that their status and information can be probed and monitored remotely.
- Quality assurance processes need to be transparent to the user despite variations between different instruments and sensors. A standard approach for detecting faulty or poor quality data early in the experiment can then be implemented.
- Triggering the download of data contained in the temporary data storage (data cache) into the permanent data storage. This is a non-trivial process, requiring the automation of metadata creation and data labeling and indexing, in large volumes, sustainably, and without human intervention.
- Security and access to the instruments and sensors. Unauthorized access could lead to tampering with the data at its source.

The DART project chose to base this group of work packages on the Common Instrument Middleware Architecture (CIMA) [3, 15]. This architecture emerged from work supported by the National Middleware Initiative specifically for connecting instruments and sensors to the Internet. CIMA allows the connection of instruments and sensors to the internet, and makes them discoverable and their results publishable using web services or the open grid services architecture. The original version of software written to CIMA was in C++. A

rewrite in Java is now being undertaken at Indiana and the University of Sydney.

3.2 Storage and Interoperability

This group of work packages related to the need to work with documents, datasets, simulations, software and dynamic knowledge representations in a secure way with controlled access. This included collection from a range of devices, secure transfer across networks, storage on high-capacity devices, management and preservation in repositories, and maintaining the integrity of the datasets.

The digital objects that DART stored needed to be managed, preserved, persistently identified, aggregated and disseminated in flexible ways. Of the available pieces of widely used repository software, Fedora has been found by a range of projects to be the best match for these requirements. Fedora is "an open source, digital object repository system using public APIs (application program interfaces) exposed as web services" [4]. Its architecture is very flexible, and provides significant advantages as a platform on which to build other applications.

In particular, in a DART context it provided the ability to store and manage complex objects and the relationships within and between complex objects. The ARROW [5] team at Monash University have been using Fedora for nearly four years now and are one of a small number of projects which are collaborating as part of the Fedora Developer Consortium.

The pre-eminent technology for working with large datasets is the Storage Resource Broker (SRB), developed at the San Diego Supercomputing Center (SDSC) [6,7]. SRB is currently being rewritten and extended as the Integrated Rule Oriented Data System (iRODS) [8].

In order to extend Fedora to work with large datasets, the DART project intended to integrate SRB with Fedora, both as a replacement storage layer for Fedora itself, and as a location for content outside a Fedora repository but managed by it.

In order to build more advanced knowledge mining services in the future, the SRB Metadata Catalogue (MCAT) was semantically augmented using the Resource Description Framework (RDF).

3.3 Content and Rights

An enormous amount of research data is currently stored within personal or private archives, either on researcher desktops or departmental/institutional servers. In these locations it is largely inaccessible to, and undiscoverable by, other researchers or the public. This group of work packages investigated methods, incentives and technologies to motivate researchers to submit their research data and results into institutional repositories. This included the development of:

- Simple user interfaces and workflows to enable researchers to easily deposit documents, research data and results into institutional repositories.
- Tools and services to enable researchers to easily select and attach standardized licenses defining access and re-use rights to their data and research results. These tools were based on the outcomes of both the Creative Commons [9] and the more recent Science Commons [10] initiatives.
- Guidelines for information management best practice in research teams, arising from embedding information professionals into such teams as research partners.

Note that assessment technologies that support qualitative and quantitative assessment of research deposited within institutional repositories also provide additional incentives for researchers to deposit their results. One example of this is the Research Quality Framework [11] being introduced in 2008 to assess the research output of Australian universities (based somewhat on the Research Assessment Exercise (RAE) in the UK). It is to be anticipated that over time document research outputs (journal articles, conference papers, etc.) will be augmented by supporting materials such as datasets and software models. Indeed, for the newer types of science these supporting materials will be necessary for the assessment of the validity of the research.

3.4 Annotation and Assessment

This group of work packages related to the tools and services that would enable peers to attach reviews, opinions, comments or assessments to research data, reports, publications etc. These annotation and assessment services can serve either as an alternative, or addition, to existing peer-review mechanisms. This can be seen as a completely new certification function made possible through this new distributed networked environment. Two annotation approaches were trialed. The first built on existing annotation research dealing with annotations that were managed and stored external to the digital

objects. The second approach built on existing work to create collaborative documents including annotations.

The first work package concentrated on extending and refining existing annotation tools to enable annotation of digital objects held within the Fedora and SRB research repositories. The second work package concentrated on tools to support collaborative annotations, thus enabling research communities to document shared practices and assessments. This involved the refinement and deployment of software called Co-Annotea. This is a derivative of the Vannotea software [12]. Vannotea is designed to enable real-time annotation of complex digital objects (images, video, 3D objects) by geographically distributed groups within a videoconferencing environment. A third work package focussed on the development of secure authenticated access to annotation servers through the development of a Shibboleth-based interface to the W3C's open source Annotea server [13], storing authorisation information as XACML. This allowed different groups who might want to annotate resources for different purposes (such as referees, grant committees, researchers) different levels of access. A fourth work package involved piloting the use of hosted wiki-style systems linked to research data repositories to facilitate interaction between researchers and research groups. In fact, the decision was made to build on top of Plone to take advantage of its superior content management capabilities.

At present, the final publication is seen as the only research record worthy of capture and curation. Both the annotation and wiki technologies described above also allowed for the capture of the records of some of the collaborative activity around the datasets and other research outputs.

3.5 Discovery and Access

This group of work packages related to tools and services that enabled researchers and readers to search, browse and discover resources within the repository and access them, either under controlled conditions or in an unrestricted way.

It involved the development of search interfaces that provided access across distributed archives implemented in SRB and Fedora. Ontologies and the semantically-augmented MCAT RDF data store were developed to provide semantic interoperability across heterogeneous metadata schemas.

In addition, one work package developed and provides access to a centralized repository/registry of metadata schemas and ontologies. Metadata schema registries enable the publication, navigation and sharing of information about metadata. This registry acts as the primary source for authoritative information about recommended metadata schemas. It enabled the sharing and re-use of existing metadata schemas and application profiles, thus enhancing interoperability and reducing costs and effort. This work package built on the open source software tools being developed within the JISC IE Metadata Schema Registry Project (IEMSR) by UKOLN and ILRT [14].

4 Outputs and Future Work

4.1 DART Outputs

The DART website [15] contains a complete set of reports, demonstrations and source code produced by the project, as well as the supporting documentation, and should be seen as the best place to locate the project outputs. Many of the workpackages have produced very comprehensive reports. Those dealing with CR1: Move data from personal to secure alternatives [16], CR4: Improve information management practice in research communities [17], and CR6: Clarifying legal issues to improve content deposits [18] are particularly worth reading.

4.2 ARCHER

The next stage of this work program is being undertaken in the Australian ResearCH Enabling enviRonment (ARCHER) project [19]. This project, funded under the same program as DART, aims to adapt the DART tools, build on existing open source projects, and commission new modules.

The intention is to produce a production-ready toolkit of software modules and a deployment environment to support the needs of e-researchers. The target for ARCHER is data-centric collaboration, with the software designed to support the research process up to the point of publication. This is in part because of the DART experience that researchers were not ready (or did not yet need) support for the complete integrated e-research lifecycle.

ARCHER is building modules to support data acquisition, storage, offline object management, collaborative workspaces, dataset deposition and export, and metadata management. This will be presented using a

Gridsphere [20] portal and integrated at the back end using Kepler [21]. The Virtual Organisation management subsystem will be integrated from the RAMP [22] project, and the authentication will rely on the Australian Access Federation (AAF) [23] which is currently being set up. A Guest Access facility will be created (in partnership with RAMP) for those users who are not in the AAF.

ARCHER has been asked to engage with elements of two of the capabilities identified by the Australian National Collaborative Research Infrastructure Strategy (NCRIS) [24] in refining its solution. These elements are non-embedded BioInformatics within the Evolving Bio-Molecular Platforms and Informatics [25] (sometimes called 5.1) capability and GeoChemistry within the Structure and Evolution of the Australian Continent [26] (5.13) capability.

The ARCHER project commenced in early 2007 and aims to complete by mid 2008.

5 Lessons learned

The project outputs document what the project achieved in greater detail than is possible in an overview paper. This section will therefore focus on the lessons learned from the process of undertaking the project, and how these are being responded to in ARCHER.

5.1 Rates of change

There is a constant need to keep up with the latest standards and developments, particularly in the Open Source world, to prevent re-inventing the wheel. One of the challenges faced in the DART project was dealing with the rate of change in this space. This is because of the level of interest in e-research and the number of projects that are developing software.

As a result, a number of the work packages needed to be modified from their original deliverables to adapt to developments outside the project. As one instance, the work package dealing with Fedora-SRB integration needed to be changed once it was discovered that the core Fedora development team had implemented a version of this themselves.

The lessons for the successor ARCHER project (see below) were to:

- ensure good relationships with those in related projects so that we will be made aware of changes as early as possible
- keep separate parts of the project decoupled as far as possible and integrated using a standards approach to reduce flow-on impacts of changes in one area.

5.2 Project management challenges

The design of the DART bid anticipated the need for strong and ongoing communication in a multi-partner complex collaborative project, and built in a considerable project management overhead.

In practice, the level of complexity in a project such as DART involved considerable overhead, as multi partner, multi researcher, multi work package objectives needed to be merged and guided. The anticipated strong project management was not always consistently applied and the central project office on occasion lost visibility of the expenditure of project funds at the other sites.

The lessons for ARCHER were to:

- improve the level of formal and informal communication between development teams
- balance local autonomy with central obligations to the funding authority.

5.3 The need for demonstrators

The original DART bid only dealt with the need to develop the individual components shown in Figure 1. While the project intended to show the value of an integrated approach, the author of the bid (also the author of this paper) inexplicably forgot to build into the work program anything to address this need. Fortunately, wiser heads prevailed soon after the project inception, and it was decided to fund a number of demonstrators.

The DART demonstrators were intended to pull together the various DART components, working in collaboration with different groups of researchers. The original plan was to select three areas that were sufficiently different and that contained researchers at each of the three DART consortium sites. The resulting areas were X-Ray Crystallography, Climate Research, and Digital History.

As it turned out, the Climate Researchers decided that their involvement with DART would distract them from time-critical research projects. The Digital History researchers were interested in mostly using the annotation features of DART. The closest demonstrator to the original DART vision was in the realm of X-Ray Crystallography. One of the leading teams at Monash University has been working with DART to control their instruments (using CIMA) and to capture the image outputs of these instruments into SRB. DART is also

working on simplifying the process of analysing the stored images using grid-enabled software tools. This work has been extremely successful, with the researchers contributing their own funds to continue it after the end of the project.

5.4 Time-poor researchers

One of the challenges for DART has been to get sufficient access to researchers to ensure that the project meets their needs. This is because all the researchers we have been dealing with are very time poor, and need to see value very quickly before they are prepared to continue the engagement. This is not intended by way of criticism – the constraints on researchers are well understood. But it does make it difficult to engage as fully as one might wish. The only solution appears to be to deliver quick wins to encourage an ongoing relationship.

The lessons for ARCHER were to:

- deliver something early on
- structure interactions with the researchers around an existing artefact where possible (see the discussion of the development methodology below)

5.5 Developing to sparse requirements

DART (and its successor ARCHER) were both projects with a short amount of total time (12-18 months). This project constraint (as well as the other demands on our research users' time) meant that the traditional development approach where detailed requirements analysis is followed by development is followed by testing (the so-called Waterfall [27] model) was not possible. Instead, development was informed by previous interactions with researchers (and previous projects) and the end users were encouraged to comment on early versions of the software. This seemed to work reasonably well.

The lesson for ARCHER was to formalise this approach by using the Scrum [28] development methodology (a form of agile development). This structures development around a series of sprints where a product owner decides which of a number of possible features should be included in the next sprint. This ensures that development is (i) based on feedback on a working artefact and (ii) oriented towards what the users want. ARCHER is also trying to integrate 'traditional' Scrum techniques with a strong approach on usability analysis. This will take the form of both heuristic usability and user-testing. This will be an interesting challenge within the tight Scrum cycle times.

6 Conclusions

The DART project has delivered the individual software components to support the original vision of an inter-connected end-to-end e-research lifecycle. The two main unmet challenges are (i) to embed these components into a national system that supports e-research and (ii) to change scholarly practice to make use of such a system.

The NCRIS Platforms for Collaboration process [29] is currently putting in place the groundwork for such a national system in Australia. The ARCHER project, building on the work of DART, hopes to contribute towards a number of components that will make up Platforms for Collaboration.

The second challenge, changing scholarly practice, is also underway. The demands of ever bigger research tasks, the need to access and validate research outputs, and a gradual generational change among researchers will all lead to more collaborative and technology-enabled ways of generating knowledge.

It is hoped that some of the experiences and outputs of the DART project can contribute in some small way to this process.

7 Acknowledgements

The author wishes to acknowledge the significant work and contributions from the DART teams based at Monash University (led by Professor David Abramson, Dr Asad Khan and the author), the University of Queensland (led by Professors Jane Hunter and Xiaofang Zhou), and James Cook University in Northern Queensland (led by Associate Professor Ian Atkinson) for their work in the conception and implementation of the DART work packages.

The author wishes particularly to acknowledge the outstanding contributions of Dr Jeff McDonell (who was the DART Project Director from January 2006 through December 2006), and Dr Robyn Polan (who was the DART Project Coordinator from January 2007 to June 2007).

The author also wishes to thank the anonymous reviewers for their very useful comments.

The DART project was funded by the Australian Commonwealth Department of Education, Science and Training. The funding was provided through the Systemic Infrastructure Initiative as part of the Commonwealth Government's *Backing Australia's Ability - An Innovation Action Plan for the Future*.

8 References

[1] Atkins, D. et al. 2003. National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure, Revolutionizing Science and Engineering through Cyber-infrastructure. Available online at http://www.communitytechnology.org/nsf_ci_report/

[2] Waters, D. 2003. Cyberinfrastructure and the Humanities. Fall Task Force Meeting of the Coalition for Networked Information. Available online at http://www.cni.org/tfms/2003b.fall/handouts/Fall2003Handouts/H-Watersplenary.doc.

[3] McMullen, R. and Chiu, K. 2005. "CIMA: Scientific Instruments as First Class Members of the Grid", Remote Access and Automation Workshop, Sydney, 2005. Abstract online at http://mmsn.chem.usyd.edu.au/events/mmsn_ws05_abstracts.html#McMullen

[4] Lagoze, C., Payette, S., Shin, E. and Wilper, C. 2005. "Fedora: An Architecture for Complex Objects and their Relationships". Submitted to International Journal of Digital Libraries: Special Issue on Complex Objects. Available online at http://www.arxiv.org/abs/cs.DL/0501012

[5] Australian Research Repositories Online to the World (ARROW) project. See http://arrow.edu.au/

[6] Moore, R. 2004a "Integrating Data and Information Management", International Supercomputer Conference, June. Available online at http://www.sdsc.edu/dice/Pubs/ISC2004.doc

[7] Moore, R. 2004b. "Evolution of Data Grid Concepts", Global Grid Forum Data Area Workshop, January. Available online at http://www.sdsc.edu/dice/Pubs/Grid-evolution.doc

[8] http://irods.sdsc.edu/index.php/Main_Page

[9] http://www.creativecommons.org/

[10] http://sciencecommons.org/

[11] http://www.dest.gov.au/sectors/research_sector/policies_issues_reviews/key_issues/research_quality_framework/

[12] R. Schroeter, J. Hunter, J. Guerin, I. Khan and M. Henderson 2006. "A Synchronous Multimedia Annotation System for Secure Collaboratories" *2nd IEEE International Conference on E-Science and Grid Computing (eScience 2006).* Amsterdam, Netherlands. December, p 41.

[13] Barstow, A, Kahan, José, Koivunen, M-R, Swick, R. "Annotea: A Generic Annotation Environment using RDF/XML", WWW10 Developers Day, Hong Kong, May 2001

[14] http://www.ukoln.ac.uk/projects/iemsr/

[15] Ian M. Atkinson, Douglas du Boulay, Clinton Chee, Kenneth Chiu, Tristan King, Donald F. McMullen, Romain Quilici, Nigel G.D. Sim, Peter Turner, Mathew Wyatt. "CIMA Based Remote Instrument and Data Access: An Extension into the Australian e-Science Environment". Proceedings of the Second IEEE International Conference on e-Science and Grid Computing (e-Science'06). Available online at http://grid.cs.binghamton.edu/projects/publications/cima-escience06/cima-escience06.pdf

[15] http://dart.edu.au/

[16] http://dart.edu.au/workpackages/cr/cr1.html

[17] http://dart.edu.au/workpackages/cr/cr4.html

[18] http://dart.edu.au/workpackages/cr/cr6.html

[19] http://archer.edu.au/

[20] http://www.gridsphere.org/

[21] http://www.kepler-project.org/

[22] http://www.melcoe.mq.edu.au/projects/RAMP/

[23] http://www.aaf.edu.au/

[24] http://www.ncris.dest.gov.au/

[25] http://www.ncris.dest.gov.au/capabilities/evolving_biomolecular_platforms_informatics.htm

[26] http://www.ncris.dest.gov.au/capabilities/structure_evolution_of_oz_continent.htm

[27] http://en.wikipedia.org/wiki/Waterfall_Model

[28] http://en.wikipedia.org/wiki/Scrum_(development)

[29] http://www.ncris.dest.gov.au/capabilities/collaborative_investment_plan_platforms.htm